CEOS ARD analysis with Open Data Cube and EASI

Earth Analytics Science Innovation
Accelerating Innovation – Enhancing Engagement

Matt Paget | CEOS LSI-VC-16, 23-25 Sept 2024

CSIRO

CSIRO

Introduction

Who am I?

- EO data and systems
  - Processing workflows and services
  - SE and east Asia projects
  - Digital capability building
  - EASI platform admin
- CSIRO representative to CEOS WGISS
- Original OpenDataCube team member
- Based in Canberra

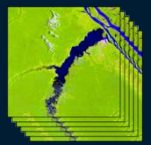
This presentation

- CSIRO EASI overview and status
- Examples
  - SAR processing and applications
  - Hyperspectral and ODC
  - Cal/Val
  - Australian AVHRR
- CEOS WGISS contributions
- Summary for LSI-VC

CSIRO

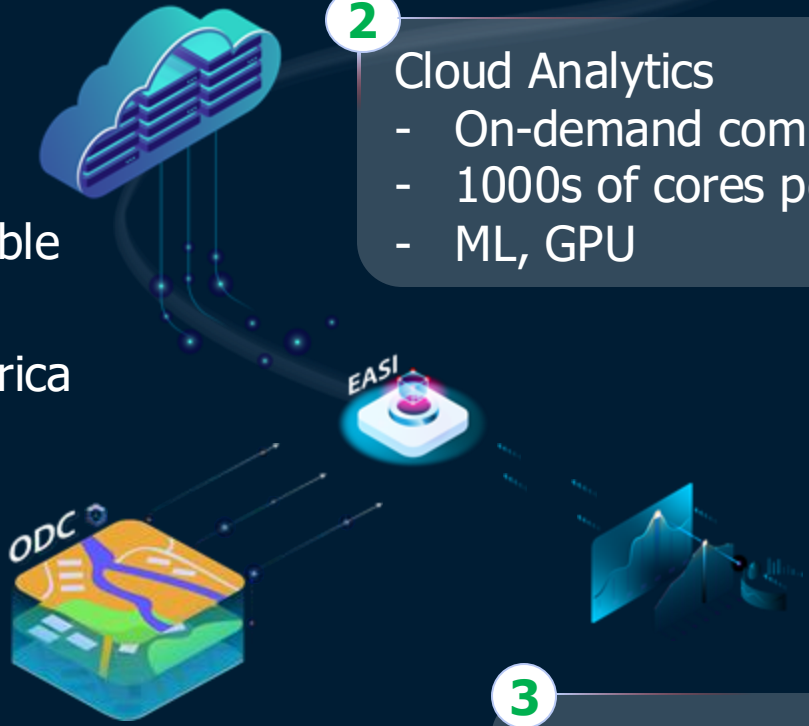
What is EASI?

- Cloud computing platform for scalable data analytics and services
- Similar infrastructure to DE Aust/Africa
- Deployed as **CEOS Analytics Lab**
- CSIRO science, partnerships and business activities



STAC SpatioTemporal Asset Catalog

0 Data sourcing, indexing, processing



1 Open Data Cube
Python data analytics

2 Cloud Analytics
- On-demand compute
- 1000s of cores per user
- ML, GPU

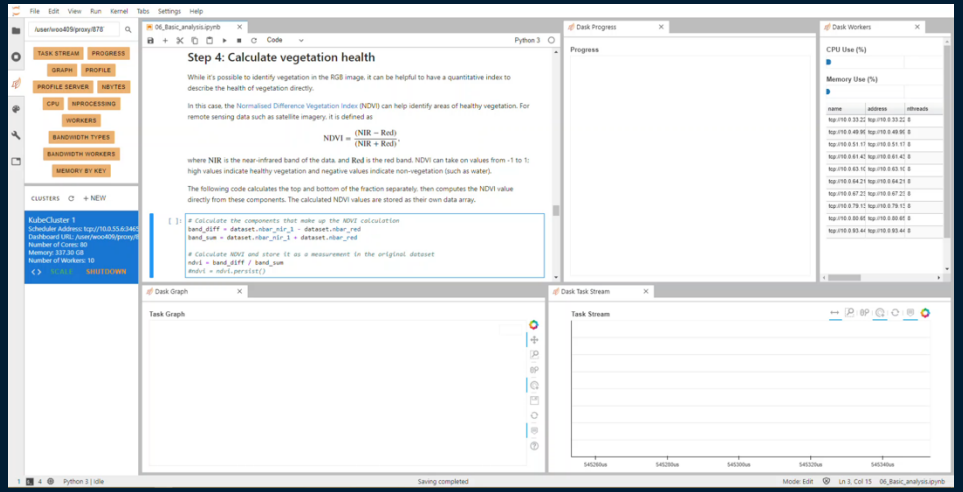
3 Analysis Ready Data
- R&D
- Science products
- Data access/services

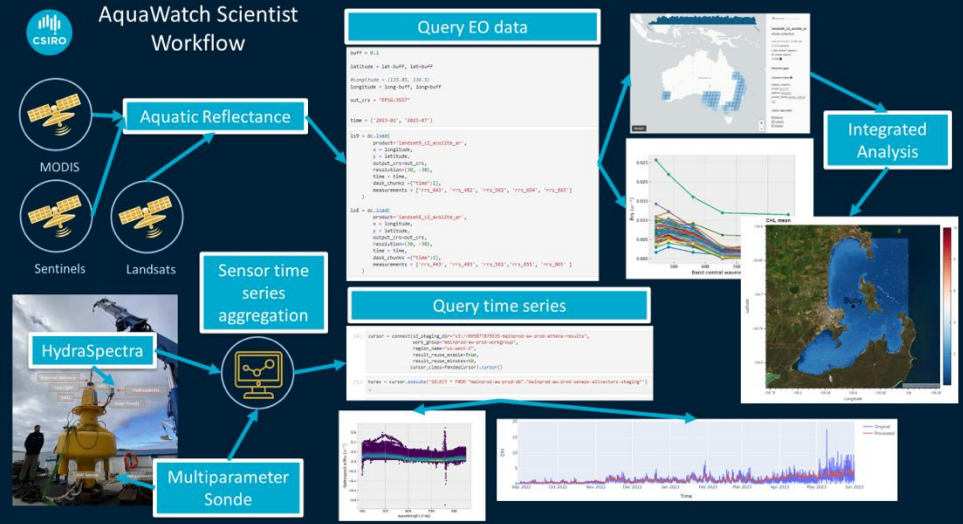
CSIRO

What is available?

- Highly scalable
  - >1000's of cores *per user*
  - *GPUs*
  - *Build your own cluster* ***on-demand***
- Multi-sensor integration
- Workflow *automation*
- Application hosting
- Enterprise and Subscription licenses for external organisations → delivering science

- Flexible, reliable, reusable
- Pay for what you use
  - Cheaper than you think!
- *Very* regular updates
- Use EASI CSIRO or Host your Own:
  - Infrastructure as Code – *setup in <4 hours!*





CSIRO Analysis Ready Data: Earth Observation

Data	Source
Digital Earth Australia	GA (AWS public)
Landsat 5,7,8,9	USGS (AWS opendata)
Sentinel-2	Element84 (AWS opendata)
Landsat / S2 blended SR	CSIRO
DEMs	NASA, SRTM, Copernicus 30 m
MODIS, S3 ocean/land	USGS, NASA
Weather and Climate	BoM Aust Global satellite/model products
Sentinel-1	Alaska SAR Facility
Sentinel-5P	Copernicus
NovaSAR	CSIRO
AVHRR	CSIRO
AVIRIS, EMIT, EnMAP, PRISMA	***Hyperspectral*** (Beta) - various suppliers: airborne & space
Himawari-8	JMA/JAXA (AWS opendata)

Plus development for Airborne, Drone, In situ…
100s managed, 1000s available

EASI Examples

Exploration Tools
- CSIRO Minerals

AquaWatch Data Services
- CSIRO Mission

AI4Missions Coastal Forecasting
- Data61 + AquaWatch

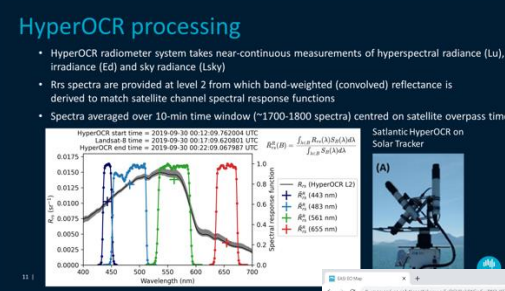
National Bushfire Intelligence Capability
- CSIRO Environment

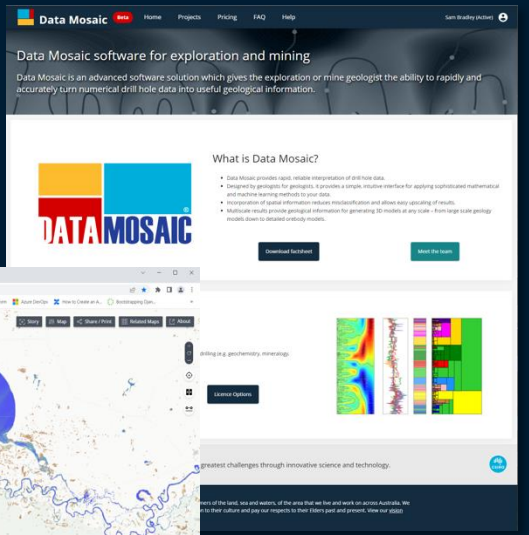
Flood Water Depth MDB
- CSIRO Environment

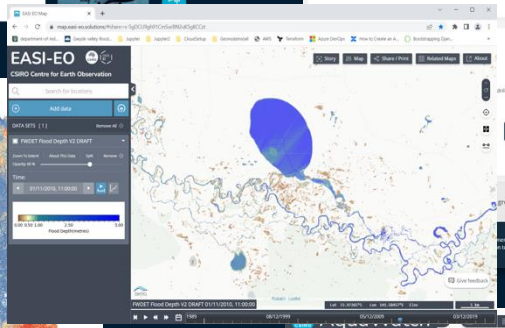
Vegetation Cover (Australia)
- CSIRO Environment

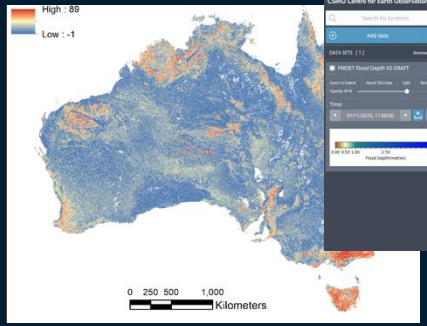
Aus CalVal…Radio Astronomy…In situ sensors

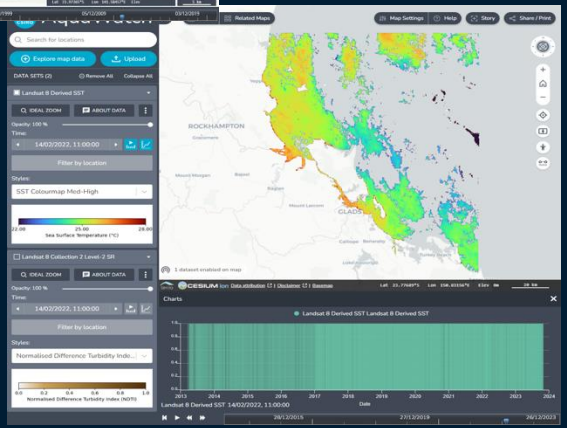
Over 300 registered CSIRO users

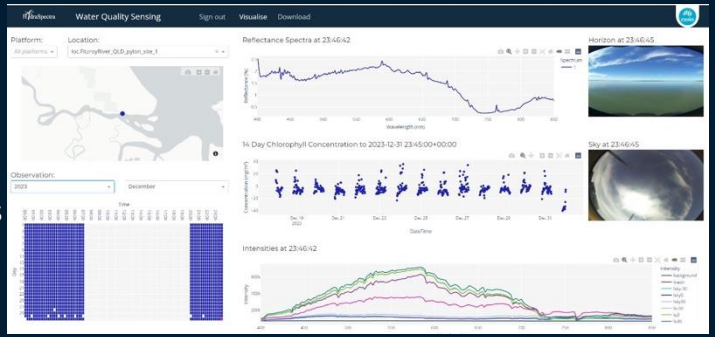








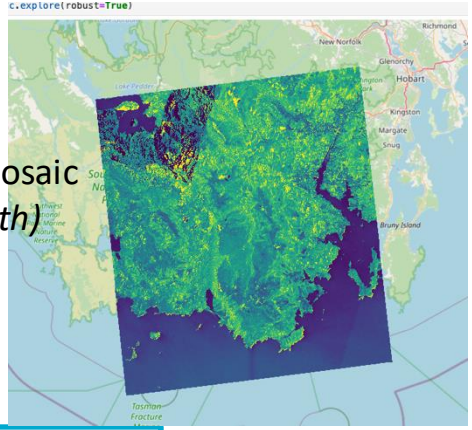




Sentinel-1 gamma0 ARD processing

Requirements	Implementation
S1 processing workflows • RTC gamma0 • Polarimetric decomposition	Workflow in **CSIRO's EASI platform** (AWS hosted) • On-demand, cost-effective • Scalable, resilient
Global coverage	Search, download from ASF
Support processing variants for R&D • Resolution, CRS, smooth/speckle • SLC, polarimetric	SNAP 10 toolbox (open license) • GPT templates (standard, customizable) • Choice of DEM (datacube or download)
Verification • CEOS ARD • Applications (notebooks, mosaics)	Cross-validation • Compare with HPC processed data • Sanity checks

Sentinel-1 gamma0 mosaic
(compliments Alex Leith)



SNAP10 gamma0 operators

1. Apply-Orbit-File
2. ThermalNoiseRemoval
3. Remove-GRD-Border-Noise
4. Calibration
5. SetNoDataValue
6. Terrain-Flattening
7. Speckle-Filter
8. Multilook
9. Terrain-Correction

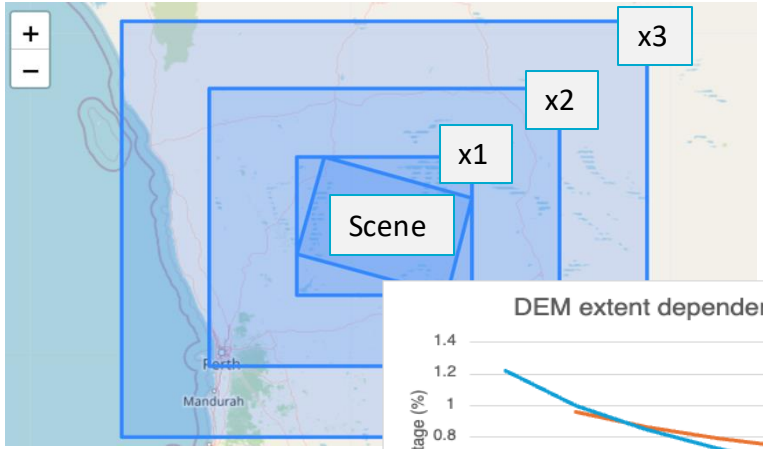
X not respected by calibration ops
Y respected by terrain ops

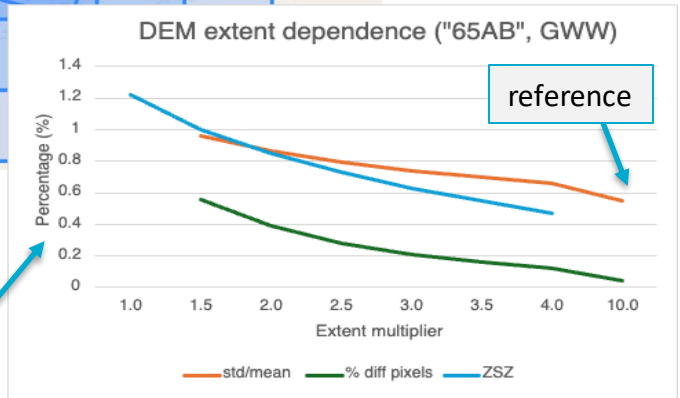
? Optional smoothing ops

S1 SNAP-10 gamma0 challenges and learnings

DEM extent (relative to S1 scene)

- Different extent DEM makes different gamma0 result
  - bilinear interpolation, or FFT?
- What DEM extent to use?
  - Physical, computational considerations





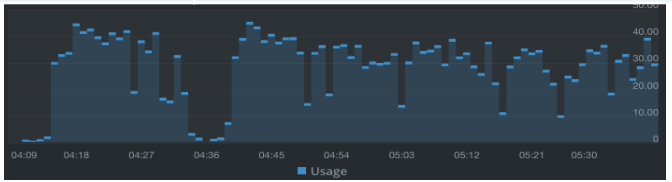
Measure of variance (low is better)

Scene processing range	
• Time • RAM	• 15 – 90 mins • 50 – 80 GB

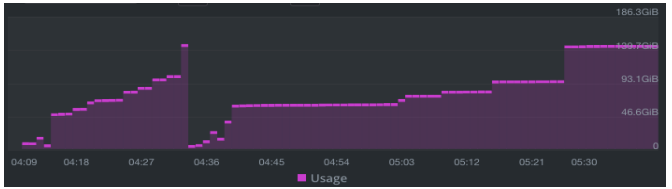
Pod specs: Cost/hour, utilisation	
• 110 Gi, 31 cpu • 60Gi, 15 cpu • 82Gi, 21 cpu	• 1x pods on 128Gib, 32vCPU • 2x pods on 128Gib, 32vCPU • 3x pods on 256Gib, 64vCPU

Workflow stability	
• Pod deadline • Allow retries	• 2x scenes / pod • 2x attempts

CPU profile



RAM profile



| First scene | ------- Second scene ------ |

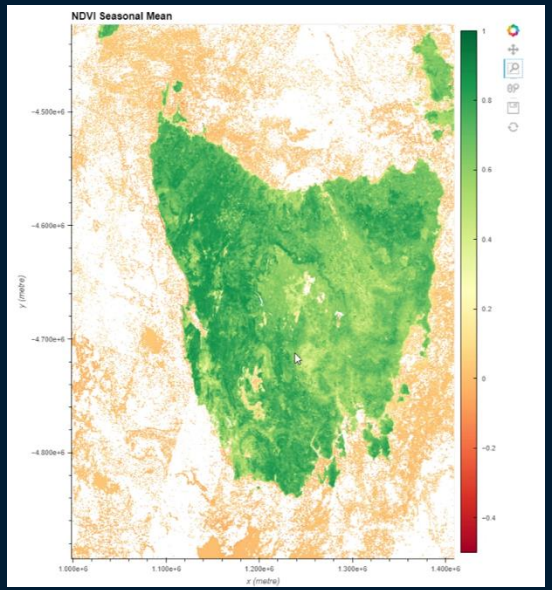
CSIRO

Multispectral (current)

- **Open Data Cube 1.8** with Dask
- `datacube.load()`
  - T, Y, X dimensions
  - 1 measurement array per band (not a dimension)
  - Bands are labelled – red, green, nir
  - Likes Cloud Optimized GeoTIFFs (COGs)
    - not so much netCDF, Zarr
  - Reprojection and data load combined
  - Dask task for every band and chunk
    - 4 spatial chunks, 1 time, 3 bands = 12 tasks
    - based on output shape only

- Supports:
  - Exploratory Data Analysis Tassie sized datasets for multi-spectral (~10 bands) for a few years of data with code like:

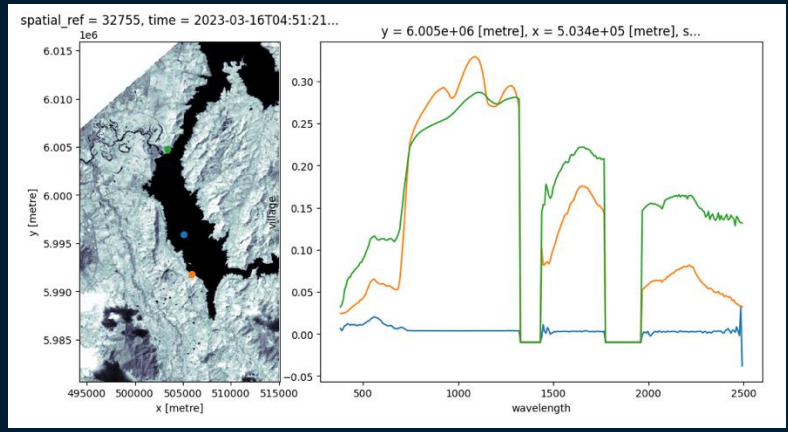
```
dataset = dc.load(…)
cloud_free_mask = masking.make_mask(dataset[qa_band], **qa_mask
cloud_free = dataset.where(cloud_free_mask)
# Calculate the components that make up the NDVI calculation
band_diff = cloud_free[ 'nir' ] - cloud_free[  'red' ]
band_sum = cloud_free[ 'nir' ] + cloud_free[  'red' ]
# Calculate NDVI
ndvi = band_diff / band_sum
ndvi.hvplot.image(groupby='season', rasterize=True)
```



CSIRO

Hyperspectral (beta)

- **Open Data Cube 1.9** with Dask
- datacube.load()
  - T, Y, X, wavelength dimensions (and options)
  - 1 measurement array for reflectance (all wavelengths)
  - Wavelength can now be sliced:
    - [0, :, :, 100:110]
  - COGs, Zarr, netCDF
  - load()
    - Native format drivers (Zarr, netCDF, …)
    - can have pre-projection hooks
  - Dask Task graph
    - more control (memory, concurrency)
    - Concurrency is product of number of sources and output shape (sources processed separately, not together)



Australian AVHRR TOA and At.Corr.

Contribution to CEOS WGISS data preservation and global AVHRR repatriation

Public access

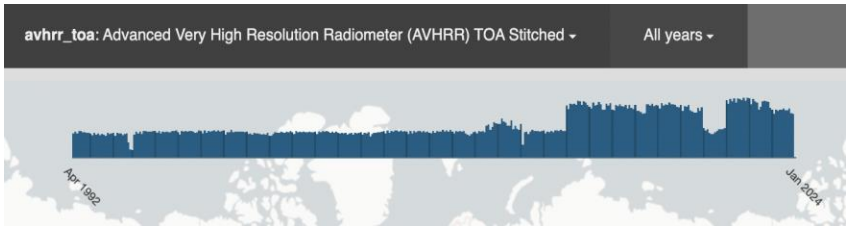
- "EASI" cloud - Scene metadata and WMS
- TERN Landscapes (forthcoming) – download

Data and analytics

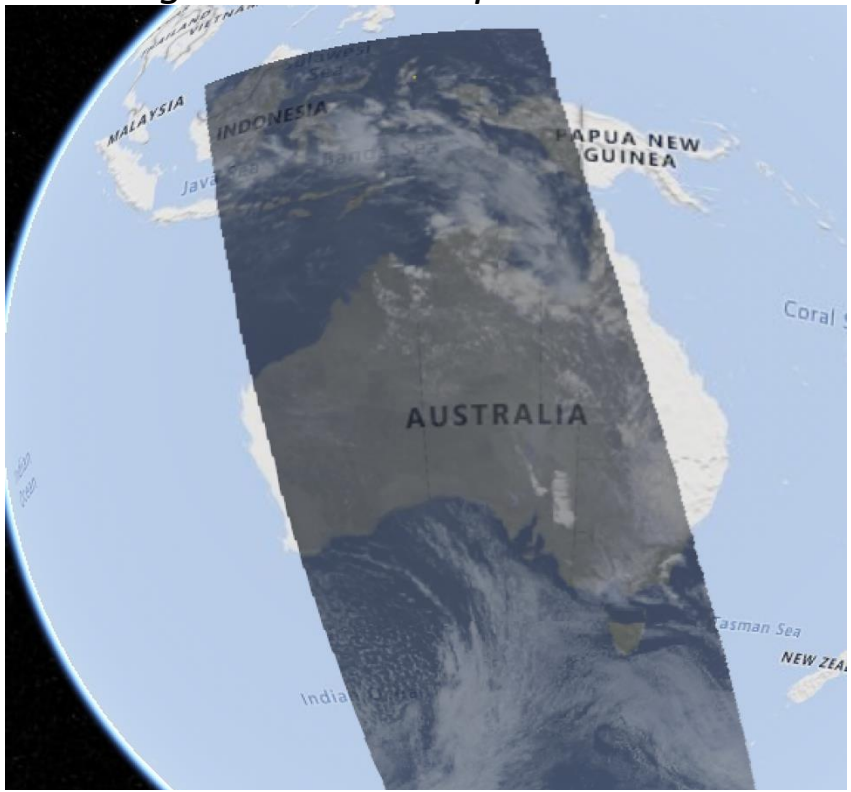
- Remapped to WGS84 COG
- "EASI" cloud – CSIRO partner login

Atmospheric correction and BRDF

- Test and upgrade at.corr. methods
- BRDF models and climatology



For more info on AVHRR on ODC contact Matt Paget (matt.paget@csiro.au)



Coverage & metadata, https://explorer.csiro.easi-eo.solutions/products/avhrr_toa
Pseudo-RGB WMS, https://map.csiro.easi-eo.solutions/#share=s-a4lVkH26vFQENph5

CSIRO

Relevant CEOS contributions

CEOS WGISS
- Interoperability handbook
  - Data and metadata services
  - Cloud best practices
- ”Middle-ware” EO data analytics and use
- Jupyter notebooks
- CEOS Analytics Lab infrastructure
- Australian AVHRR

CEOS WGCV
- Cal/val techniques, data and networks

Ecosystem extent mapping
- Developing and proving data integration and analysis

CEOS SIT and management
- Coordination, advocacy and planning

Data preparation considerations

Key points
- Cloud storage is relatively expensive
- On-demand processing is relatively cost effective

Cloud friendly storage and format
- Other Organisation manages the storage
- We prepare the metadata for ODC indexing
  - STAC metadata if available
  - Else, prepare our own metadata

Not cloud-friendly storage and format
- We download, convert to COG, store
  - Potential combination of on-prem and cloud
  - Optional processing to “ARD”
- We prepare the metadata

Thank you

Centre for Earth Observation
CSIRO Environment

Matt Paget
Matt.Paget@csiro.au

Centre for Earth Observation
CSIRO Space & Astronomy

Dr Matt Garthwaite
Matt.Garthwaite@csiro.au

Centre for Earth Observation
CSIRO Space & Astronomy

Dr Robert Woodcock
Robert.Woodcock@csiro.au

Centre for Earth Observation
CSIRO Space & Astronomy

Dr Shaun Levick
Shaun.Levick@csiro.au